Working Papers of the Priority Programme 1859
Experience and Expectation.
Historical Foundations of Economic Behaviour
Edited by Alexander Nützenadel und Jochen Streb

No 19 (2020, July)

Foltas, Alexander

Testing Investment Forecast Efficiency with Textual Data

Arbeitspapiere des Schwerpunktprogramms 1859 der Deutschen Forschungsgemeinschaft
„Erfahrung und Erwartung. Historische Grundlagen ökonomischen Handelns“ /
Working Papers of the German Research Foundation's Priority Programme 1859
"Experience and Expectation. Historical Foundations of Economic Behaviour"

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Published in co-operation with the documentation and publication service of the Humboldt University, Berlin (https://edoc.hu-berlin.de).

ISSN: 2510-053X

Redaktion: Alexander Nützenadel, Jochen Streb, Ingo Köhler

V.i.S.d.P.: Alexander Nützenadel, Jochen Streb

SPP 1859 "Erfahrung und Erwartung. Historische Grundlagen ökonomischen Handelns"
Sitz der Geschäftsführung:
Humboldt-Universität
Friedrichstr. 191-193, 10117 Berlin
Tel: 0049-30-2093-70615, Fax: 0049-30-2093-70644
Web: https://www.experience-expectation.de
Koordinatoren: Alexander Nützenadel, Jochen Streb
Assistent der Koordinatoren: Ingo Köhler

Recommended citation:

Foltas, Alexander (2020): *Testing Investment Forecast Efficiency with Textual Data.* Working Papers of the Priority Programme 1859 "Experience and Expectation. Historical Foundations of Economic Behaviour" No 19 (July), Berlin

Testing Investment Forecast Efficiency with Textual Data

Alexander Foltas

Helmut Schmidt University,
E-Mail: foltasa@hsu-hh.de

July 29, 2020

Abstract: I use textual data to model German professional macroeconomic forecasters' information sets and use machine-learning techniques to analyze the efficiency of forecasts. To this end, I extract information from forecast reports using a combination of topic models and word embeddings. I then use this information and traditional macroeconomic predictors to study the efficiency of investment forecasts.

JEL Classification: C53, E27, E22

Keywords: Forecast Efficiency, Investment, Random Forest, Topic Modeling

1 Introduction

During the past decade, new technologies made the processing of vast quantities of textual data possible and led to an increased usage of text-based empirical research in economic and social sciences. The idea of this relatively new research area is to decode the information in human communication and complement the data used in traditional research. Computational text analysis is used, for example, in measuring the political slant of media content as in Gentzkow and Shapiro (2010), where newspaper data is used to measure the slant of 433 US daily newspapers or as in Greenstein et al. (2017), who examined bias and different behaviors among Wikipedia contributors. Other authors analyzing the effect of the sentiment from Central Bank communication on fluctuations in Treasury securities (Lucca and Trebbi 2009), developed an indicator of economic policy uncertainty based on newspapers (Baker et al. 2016) or measure political risk at the firm level by investigating quarterly earnings call transcripts (Hassan et al. 2019).
Multiple researchers used textual information for economic forecasts. Predictive power for the stock market prices are found for the sentiment of the Wall Street Journal (Tetlock 2007) or through the mood of Twitter messages (Bollen et al. 2011). Beckers et al. (2017) examine the predictive ability of media-based sentiment indicators for the German inflation. Choi and Varian Hal (2012) and Scott and Varian (2014, 2015) implemented textual data in the estimation of the current level of macroeconomic variables ("nowcasting"). They use the frequency of Google search terms aggregated by week and geographic location to improve forecast regional retail sales, new housing starts and tourism activity significantly[1].
Textual analysis proofed to capture information non-observable by traditional statistical methods of measurement. As shown in the survey from Döpke et al. (2019), professional forecasters exploit non-quantified information in the forecast process. The goal of this study is to reconstruct the forecasters' set of information using textual German macroeconomic forecast reports. Subsequently, I examine whether forecasters implement this information efficiently into their prognoses. Thus, my research goal relates to multiple strands of literature: My first link are the "narrative" methods in macroeconomics, for instance, the content analysis of textual representation of business cycle expectations (e.g. Mathy and H. Stekler 2017; H. Stekler and Symington 2016). I relate to Fritsche and Puckelwald (2018), who used forecast reports of German forecasters from 1990 to 2017 in dictionary-based sentiment analysis, latent Dirichlet allocation (LDA) model and structural topic model (STM). My methodology differs as I use a combination of LDA with word embeddings, which allows me to cover a more extended period.
Furthermore, I link to the forecast efficiency literature, in particular to studies that investigate the forecast efficiency of German economic research institutes with a nonparametric approach. Forecasters have to minimize the loss function to all of their available information for achieving strongly efficient forecasts (Nordhaus 1987). Behrens et al. (2018b, 2020) examine the efficiency of growth and inflation for German forecasters with a nonparametric tree-based method. Behrens et al. (2018a) and Behrens (2019) use a

[1]Further applications may be found in the recent survey of Gentzkow et al. (2019).

joint efficiency analysis of German trade forecasts in a multivariate setting. As this reasonable novel approach allows a high number of explanatory variables, I use it to test if the textual data issued by forecasters is fully incorporated in their forecasts. To my knowledge, forecast efficiency has not been tested for German investment forecasts yet. As the crucial role of investments for economic growth, my goal is to fill this gap in research. For this purpose, I use traditional indicators together with textual information and prove whether I can reject strong forecast efficiency.
I present my utilized topic model and the textual data in Sections 2 and 3. In Section 4, I investigate the results of my textual analysis. Sections 5 and 6 present the random forest method and the forecast data. Section 7 summarizes the results of my forecast efficiency test and Section 8 contains my conclusion.

2 Topic Models with Word Embeddings

Topic models are hierarchical probabilistic models developed for the automatic analyzation of text corpora. The goal is to use machine learning techniques to discover semantic patterns that reflect underlying topics which got combined to form the document. The most basic topic model is the latent Dirichlet allocation (LDA) (Blei et al. 2003). The idea of LDA is that each document contains a distribution over latent topics, which contain a distribution over words. The respective topic proportions provide a low-dimensional representation of the content of each document.
Although the basic LDA is a powerful model, it faces some shortcomings, which made it an insufficient tool for the analyzed corpora. Firstly, its quality suffers from large sizes of vocabulary. On large collections of documents, it is required to severely prune the vocabulary to fit interpretable topic models. Typically, the researcher removes the most and least frequent words of the vocabulary and thus faces the risk of removing important terms, and therefore, limiting the scope of the models (Dieng et al. 2019).
Secondly, LDA is designed for categorical data and faces problems with the long period of the sample (Blei and Lafferty 2006). Therefore, it is not capable of handling the evolution of topics and written language. For instance, the reference value for economic growth was the gross national product ("Bruttosozialprodukt") until 1992, when it was changed to the gross domestic product ("Bruttoinlandsprodukt"). Both terms were frequently used in sections analyzing the growth, and their prevalence values for different topics should be highly positive correlated as both terms are used in the same context. As there are only very few documents in which both terms have a high co-occurrence, standard LDA would result in a highly negative correlation. A widespread solution is to limit the time frame of the used documents. As my goal is to use the topic proportions for time series analysis, I cannot use this workaround.
To solve these problems, I combine LDA with word embedding, which is a method of mapping words in vector space and thus representing their meaning (Panigrahi et al. 2019). With this method, terms used in similar contexts like "Bruttosozialprodukt" and "Bruttoinlandsprodukt" get placed nearby in vector space, although they seldomly occur

nearby in the documents. This method results in a correct classification of both terms as highly positive correlated and furthermore, it does not suffer from a high number of tokens and is resistant to stop words (Dieng et al. 2019).
Let D express a number of documents which get merged into a text corpus with $V = \{w_1, w_2, ..., w_V\}$ unique tokens, then C signifies the word-word co-occurrence matrix with $V \times V$ dimensions. The element C_{ij} denotes the number of times the context word c_j has occurred in the near of w_i, which means how often it is counted in an n size window centered around the w_i.
Column C_w of the matrix C may be interpreted as a *document* formed from the counted tokens around the centered word w (Panigrahi et al. 2019). For the creation of these, I suppose a generative probabilistic model, where the *documents* are characterized by random distributions $\theta_{1:V}$ over K latent topics. The topics are multinominal distributions $\beta_{1:K}$ over the vocabulary. Each *document* C_w of the co-occurrence matrix C is supposed to be generated as follows:

1. Draw topic proportions $\theta_w \sim Dirchlet(\alpha)$.
2. For each context word $c_{w,n}$ in C_w:
   a) Draw topic $z_{w,n} \sim Multinominal(\theta_w)$.
   b) Draw token $c_{w,n} \sim Multinominal(\beta_k)$.

The variable $z_{w,n}$ describes the topic assignment for the nth word in the wth *document*. The complete model can be formulated as a joint distribution of hidden and observed variables (see also Blei 2012):

$$p(\theta, z, c, \beta|\alpha, \eta) = \left(\sum_{i=1}^{K} p(\beta_i|\eta)\right)\left(\sum_{w=1}^{V} p(\theta_w|\alpha)\sum_{n=1}^{N} p(z_{w,n}|\theta_w)p(C_{d,n}|z_{w,n}, \beta)\right), \quad (1)$$

with α as Dirichlet prior for the topic-document distribution and η as Dirichlet prior for the term-topic distribution. The Dirichlet distributions are the conjugate prior of the multinomial distributions. The parameter α and η are K and V dimensional vectors which determine the likelihood of each possible probability distribution. As Equation 1 is analytically intractable, it is estimated most commonly with collapsed Gibbs sampling (Griffiths and Steyvers 2004; Heinrich 2009).
To obtain the topic proportions $\theta_{1:D}$ for each document D, I implement the distribution over latent words for each topic β_k into a standard LDA model. Instead of merging all documents into a co-occurrence matrix with $V \times V$ dimensions, the standard LDA model uses a document-term matrix with $D \times V$ dimensions. Both approaches are visualized in the figures 1 and 2. The word embedding approach uses a distribution over topic proportions α_1 and draw a topic proportion for each *document* C_w. Then for each context word c a topic gets drawn based on θ_w, and subsequent the token gets drawn based on the distribution over latent words β_k. The LDA model uses a different distribution over topic proportions α_2 and draws a topic proportion for each document D of the text corpus. Here for each word w in the document a topic gets drawn based on the topic proportions θ_d and the distribution over latent words of each topic β_k. Through obtaining the distributions $\beta_{1:K}$, I use in vector space embedded words inside an LDA model. Thus I can identify tokens as part of the same topic based on their meaning, which significantly improves my model and allows me to use an extended period for my model.

3 The Corpus

The investigated corpus consists of business cycle forecasts for Germany from four German economic research institutes (DIW, ifo, IfW, HWWA) and the "joint diagnosis" of the five/six leading research institutes. The publication dates of the 584 documents range from 1960 until 2017, with varying publishing frequency. While the newer publications can be found online, the majority of the documents had to be obtained by scanning and using Optical Character Recognition (OCR). Although I manually checked all digitalized text pages for scanning or OCR errors, some words still might be incorrectly identified.

Usually, the complete forecast reports include international parts, policy advice or methodical explanations, which is not part of my research interest. The texts were manually fragmented, and only the description of the recent economic development and the textual expressions of forecasts for the German economic development are used.
To make the corpus suitable for topic modeling, I applied multiple pre-processing steps to the original texts. Firstly all numbers and all punctuation were removed from the documents. Also, all letters were converted to their lower case. It is not required to remove stopwords without any content-related meaning because they should be gathered into stopword topics by my used method. Nevertheless, we decided to remove them to minimize the number of stopword topics. Therefore I used the German stop word list included in the tm R package (Feinerer 2013; R Core Team 2019) and extended them with self-identified stop words. These are mainly words related to the structure of tables and figures, date-related expressions, and names of institutions and their publications as well as the most common words without any economic meaning. Most authors remove the most and the least common terms as well, based on the times they occur because LDA achieves better results with smaller sizes of vocabularies. As I do not suffer from this problem, I did not apply this procedure.
Following Dieng et al. (2019), I use the Topic Quality to measure the optimal value for K. Topic Quality is measured by multiplying the Topic Coherence and the Topic Diversity. Topic Coherence provides a quantitative measurement for the interpretability of a topic (Mimno et al. 2011). It is obtained by approximating the mutual information of the ten most likely words of a topic. The Topic Diversity is defined by the percentage of unique words in the top 25 of all topics. Topic diversity of 0 means all words in the top 25 of a topic also occur in the top 25 of a different topic. Thus a diversity of 0 suggests redundant topics, while a diversity of 1 indicates highly varied topics. As figure 3 suggests I achieve the highest Topic Quality at $K = 24$.

4 Results of the Topic Models

Figures 4 and 5 show the topic proportions of the 12 most and the 12 least common topics. I find a clear trend towards more diversity of the topics over time. Remarkably, five topics have a combined topic proportion of up to 0.75 in the 1960s, with a proportion of up to 0.25 for Topic 5 alone. With the beginning of the 1990s, the forecast reports show a developing diversity. This could be explained partly with a growing length and frequency of the publications, which enables more specialized topics.
Table 1 shows the labels I granted each topic based on the top terms. Although I show only the first ten top terms, I took a more in-depth look for the topic identification. Six topics are gray underlined, as they cannot be assigned to a specific economic subject and thus get excluded from further analysis. Topic 3 is a small topic that mainly includes the names of different forecast reports and institutes, which mistakenly did not get excluded during the preprocessing. Topic 11 focuses on the underlying statistical methods, which are included in newer publications. Topic 9 is an overview topic, which could be found in

the introduction or conclusion of a forecast report. This is indicated by the topic diversity of 0, which means that all top terms also occur as top terms in other topics. Topic 13 includes mainly general words that can be used in conjunction with any economic subject such as "expand", "abundant" or "relevant".
I also classify Topic 1 and 5 in the context word category. Although the first topic could be assigned as a business expectations and the second into an international climate topic, there are reasons against such a classification. An in-depth inspection shows that the overwhelming majority of words are context words that do not have an economic meaning on their own. Furthermore, these topics include a high number of conjunctions and verbs. Especially during the early periods, they are vast in size, with a combined topic proportion of 0.42 in 1960. In opposite to most other topics, they include a high number of different terms, each one with a relatively low frequency. For example, the top term of Topic 5 (12,3% of all tokens) occurs approximately 650 times within the topic. On the other hand, the top term of Topic 24 (1,4% of all tokens) occurs over 6000 times. Therefore, I regard them as a pool for words that could not be assigned to a specific economic subject. I conclude that the earlier forecast reports had less room for specific subjects because of their limited size of only a few pages. The growing extent of the forecast reports allowed sections about specific issues, which causes a shrinking to a combined topic proportion of about 0.11 of the context word category.
The most common, specific economic subject, Topic 2, centers around different forms of investment with "building investment", "equipment investments" and "investment" in the top terms. The proportion of the topic shows a clear downward trend, which is partly offset by the growing proportion of Topic 23, which focuses on investments as well. As there are different emphases of both investment topics, the consideration of both topics shows a shifting perspective on investments. The top terms of Topic 2 "demand", "production" and "cyclical" strongly associate investments with the business cycle. This topic was especially popular before the 1980s when the Keynesian paradigm was prevailing in the economic theory. Nowadays, the connection between macroeconomic investments and the economic cycle seems to be less emphasized. Instead, investments seem to be more associated with the improvement of the infrastructure and, thus, with longer-term growth prospects. The addition of both topic proportions show that the relevance of investments varied overtime around approximately 0.15.
Topic 16 contains "gross national product" as the leading term, it appears more than twice as often as the second most frequent word, and a series of terms related to the size or the measurement of the GDP, for example, "increase", "growth" and "seasonally adjusted". The popularity of the GDP topic develops similar to a quadratic function. During the 1960s and after the 2000s, the topic takes a considerable portion of the forecast reports, with a phase of 30 years of relatively less popularity.
Topic 14 and Topic 20 capture economic policy as I conclude by the top terms "financial policy", "monetary policy", "measures" and "economic policy". Despite the considerable similarities between these two topics, there are differences in the focuses. Topic 14 links the economic policy with "federal government", "decision" and "resolved", and therefore emphasizes the decision process. In contrast, Topic 20 includes terms connected to the impact of those policies, for example, "effect", "upswing" and "corporations". Both top-

ics reach their maximum height around 1980 with a combined topic proportions of 0.19, which is reduced to only 0.04 in 2017.
This development is partially offset by a growing proportion of more specialized policy topics. For instance, Topic 4, which centers around government spending, budget deficit and financial policy, nearly doubled its topic proportion. The proportion fluctuated around 0.015 until the 2000s, where it leaps to a new center around 0.025. Notable is a peak in 2010 and 2011 at approximately 0.04, during the peak of the euro crisis and the implementation of the balanced budget amendment. Another growing subject is Topic 22, which regards taxes and social insurances. The increase is implemented stepwise. Until the mid-1960s, the proportion is near zero. Then, it centers around 0.01 until the end of the 70s. Afterward, the proportion fluctuates around 0.025 until the middle of the 90s, when it leaps to the new mean at approximately 0.05. Remarkable outliers are in 1968 and from 2004 to 2007.
Topic 17 focuses on monetary policy, with "Bundesbank", "interest" and "money supply" as top terms, but without any price- or inflation-related terms. Those are gathered in Topic 6 and combined with the terms "oil price", "exchange rate" and "unit labor costs". Due to the strict separation of monetary and price-related terms, I conclude that forecasters analyze both topics independently, despite their apparent economic connection. Both topic proportions have a stationary course around a mean value, with outliers in which the topics are of a higher relevance. For Topic 17, the mean is at 0.025 and outliers are found in 1971, 1994, 1995, 2008-2013. The mean of Topic 6 is at 0.03 with positive outliers in 1969, 1973, 1986, 2002-2005 and 2008. The variance of the proportion of Topic 6 is overall higher than in Topic 17.
Another important topic is Topic 10, which could best be described as a sectoral analysis. The top words are "business" and "industry", while the less essential terms are referring to specific branches. Examples are "services", "retail", "trade", "construction" and "sector". While the topic is quite substantial in the 1960s with proportions of around 0.12, the releases focus nowadays much less on this issue. The mean proportion of 0.018 since 1980 shows the rapid shift away from the sectoral analysis of the economy and therefore shows the different character of the early forecast reports.
There are three labor market-related topics. The most frequent one of these, Topic 8, focuses on wages with terms as "wages", "tariff", "standard wage" and "profit". Topic 18 include employment-related terms, for instance "persons", "unemployment", "employed", "work time" and "working population". Topic 7 combines employment-, unemployment- and immigration-related terms, for instance "persons", "unemployment benefit" and "refugee". The topic proportions of the three topics combined fluctuate around 0.05 until the mid-1980s. Afterward, it shows a positive trend with peaks of around 0.15 and 0.16 in 2003 and 2014. The proportions of Topic 8 and especially Topic 7 rise over time, while Topic 18 declines in recent years.
There are two trade-related topics. The bigger one, Topic 15, includes "German", "export", "competitiveness", "euro area" and, but much less frequent, "import". The topic proportion shows an upward trend and the increasing emphasis on exports as the German growth strategy. The second one, Topic 19, focuses more on the balances with "current account", "deficit", "account", "surplus", "term" and "trade" as top terms. Another

subject is the medium-sized recession Topic 12. Interestingly, it combines the terms "recession", "financial crisis" and "uncertainty" with the terms "winter" and "weather conditions". Two smaller subjects are Topic 21, which centers around business expectations, and Topic 24, which focuses on household income and consumption.
It is not possible to thoroughly compare our topic models to the structural topic models of Fritsche and Puckelwald (2018), as they published only four of their forty topics. The topic proportions of the shown topics are most of the time at around zero with high positive values for a limited period. As the authors use more topics while regarding only half of the time span of this paper, I assume that their topics show rather different times of economic discourse than economic issues. Therefore, the results of the structural topic model analysis of the investigated corpus are harder to interpret and not suitable for a forecast efficiency test.

5 Random Forest

The most common approach for testing of the forecast efficiency is to estimate a regression of the form $e_{t+1} = \alpha + \mathbf{X}_t\beta + u_{t+1}$, where e_{t+1} is a series of forecast errors, $\mathbf{X}_t$ describes the set of predictors, and u_{t+1} the error component. As Behrens et al. (2018a) describe, there are some serious problems using regressions for testing the forecast efficiency. Most important is the limitation through the degrees of freedom, as I use up to 37 predictors with an N ranging from 24 to 48. A possible solution to this approach is to use a subsample of $\mathbf{X}_t$ for the estimation, but this seems impractical, because of the large numbers of estimators, and raises questions of how to nonarbitrary choose from a large number of possible subsets. A second potential problem of the standard approach is that it captures nonlinear links between e_{t+1} and x_t only in a rudimentary way. Nonlinear links could occur if forecast errors vary with the state of the economy (Sinclair and H. O. Stekler 2013) or if interaction effects between predictors have predictive value for the forecast error. Because I expect the content of the forecast reports to be influenced by the economic indicators, interaction effects are considered as likely. I prefer the tree-based method of Behrens et al. (2018a) for my data, as it leads to considerably less effort than testing all conceivable interactions of my 37 predictors in each of my 36 time series. Furthermore, random forest provides potentially more new insights if unexpected or more complex interactions exist and is much less likely to suffer from omitted-variable bias.
I use a univariate regression tree of the format $e_{t+1} = T(\mathbf{X}_t)$ to estimate my model, whereby T denotes the regression tree. Tree-based models are nonparametric approaches that split the predictor set in nonoverlapping regions representing a relatively homogeneous outcome of the response variable. The regions are created by applying binary hierarchical recursive splitting rules to partition the predictor set. The tree consists of a root, interior nodes and terminal nodes. At each node N a splitting function Φ is applied to partitioning the predictor set into one left N_L, for $x_t < c$, and one right node child node, for $x_t \geq c$. Thereby the partitioning predictor x_t and the splitting point c have to

be chosen to minimize the node impurity measure, defined as:

$$SS(N_j) = \sum_{t \in N_j} (e_{t+1} - \bar{e}_{N_j})^2, \tag{2}$$

where $j = R, L$ and the summation is over the data sent to corresponding nodes N_j. Node impurity is calculated by computing the sum of squared differences between the forecast errors sent to node N_j and their region-specific mean $e_{\bar{N}_j}$. The splitting function Φ uses the node impurity measures for N_L and N_R and thus controls for homogeneity at both nodes:

$$\Phi(S, N) = SS(N) - SS(N_L) - SS(N_R). \tag{3}$$

The goal is to choose a split out of all possible splits S so that the splitting rule maximizes the reduction of node impurity $\Phi(S, N)$ at each splitting point. After identifying the optimal splitting point, the same method is applied for the next hierarchical level of the regression tree. This process is continued until the forecast errors at the terminal node reach some predefined minimum or the number of nodes reaches a predefined maximum.

An example of such a regression tree is shown in figure 7. The splitting function chooses x_1 as partitioning predictor for the top level with the splitting point c_1. The algorithm chooses a second split based on the predictor x_2 with the splitting point c_2 for the data sent to the left branch of the tree. If the condition $x_2 \geq c_2$ is met, the final leaf is reached with $\bar{e}_3$ as the mean response of the forecast errors. Otherwise, the data gets divided again with the predictor x_1 and the splitting point c_4. After the third level split, the tree predicts mean forecast errors of $\bar{e}_1$ for the left branch and $\bar{e}_2$ for the right branch. Analogous to this procedure, the predicted mean forecast errors for the case of $x_1 \geq c_1$ are $\bar{e}_4$, $\bar{e}_5$ and $\bar{e}_6$.

A single regression tree is, because of its hierarchical structure, a high-variance predictor. I use a random forest model, which grows a high number of independent trees to solve this problem. Therefore a large number of random data samples are created with bootstraps. A random tree is created for every data set, which uses only a random subset of the predictors for every partition. This proceeding leads to a decorrelation of the predictions from the individual trees and a much lower variance of the random forest in comparison to individual regression trees. The random forest method offers several benefits in comparison to other machine learning techniques. It works well with irrelevant data inputs and is robust to outliers. As it handles missing values and mixed data types, no time-consuming data preprocessing is needed. Additionally, tree-based methods are computationally faster than other machine learning techniques, particularly in the case of high numbers of observations. Furthermore, the results have reasonable interpretability, especially in comparison to neural nets or support vector machines. Due to these reasons, random forests have become the "off-the-shell" method, which means data scientists often apply them as the first method in data mining. (Hastie et al. 2009, chapter 10). One of the biggest disadvantages of this method is lower predictive power in comparison to other machine learning methods. I consider this drawback as outweighed by the higher robustness and interpretability.

As only about two-thirds of the bootstrapped data is used to fit the random forest, I use the withheld out-of-bag data to analyze the performance of a random forest in an artificial out-of-sample experiment (see James et al. 2013, chapter 8.7). My algorithm follows Behrens et al. 2018a (with B_F as the number of bootstrapped samples):

For $(b = 1 : B_F)\{$

1. Estimate a regression tree, T_b, on every bootstrapped data, b. To this end, recursively repeat for every split the following steps:
   a) Select m variables at random from the predictors in $\mathbf{X}_t$).
   b) Given b and m, use the split function defined in Equation 3 to identify the best splitting variable and the best split point.
   c) Form the left-hand side and the right hand side of a node.
2. When the terminal splits have been computed, store the estimated tree, T_b.

$\}$

---

6 The Data

I analyze the forecasts of gross fixed capital formation (CAP), Construction (BUILD) and investment in machinery and equipment (EQIP), by three German research institutes (DIW, ifo and IfW) from 1970 to 2017. I do not investigate the forecasts of the HWWA and the "joint diagnosis" as they did not have predictions for the entire period. Nevertheless, I used their forecast reports to create topic models to capture the economic discourse more comprehensively. As I am not interested in the textual information of a specific document, but rather in the discourse of economic institutes as a whole, a broad textual corpus is necessary. Therefore, I use a larger number of documents to create our topic model, which offers me a more accurate representation of relevant economic issues. As forecast reports are only included in the efficiency tests if they are associated with an investigated prognosis, the additional documents do not affect the forecast efficiency tests.

The publication frequency of the regarded forecasts varies over time and institute. I concentrate my analysis on four forecast horizons: forecasts of the fourth quarter for the same year (H0), of the second quarter for the same year (H2), of the fourth quarter for the next year (H4) and of the second quarter for the next year (H6). I define forecast errors as: $e_{t(h),i,j} = \hat{y}_{t(h),i,j} - y_{t,j}$, where $e_{t(h),i,j}$ expresses the forecast error in year $t = \{1970, ..., 2017\}$, at forecast horizon $h = \{0, 2, 4, 6\}$, made by institute i, for aggregate j. The forecast for period t is denoted by $\hat{y}$ and y denotes the realized value. I use first release data from the German statistical office to receive values for CAP, BUILD and EQIP.

Table 2 shows the descriptive statistics of my data set. The number of observations varies between the different macroeconomic aggregates from 24 to 49 for CAP, 36 to 49 for BUILD and 30 to 46 for EQIP.

After testing the impact of the topics on the forecast error, I use traditional indicators to compare their performance with the topics and additionally search for possible interactions. I follow Behrens et al. (2018a,b) and use several macroeconomic variables as a set of predictors, which are presented in Table 3. In order to model the information set which was available to a forecaster when a given forecast was formed, I account for different publication lags of the predictors based on research by Drechsel and Scheufele (2012).

7 Empirical Findings

I use the R packages "randomForest" (Liaw and Wiener 2002) to compute the random trees and "rfUtilities" (Evans and Murphy 2018) to estimate permutation tests. As Probst et al. (2018) have shown, random forest works well with the default settings and is far less tunable than other algorithms, wherefore they are used in my baseline settings. I define *nodesize*, the minimal size of a terminal node, as 5 and *mtry*, the number of candidate variables randomly selected for each split out of the predictor pool, as $p/3$, with p equals the total number of predictors. The values of these hyperparameters are varied in robustness checks. As Probst and Boulesteix (2018) have theoretically proven, more trees always improve the performance of random forest. Therefore, I set the number of trees to 1000, which is regarded as a sufficiently high number.
Table 4 shows the results of my baseline specifications. Bold numbers indicate significance at the 10% level. In the four first columns, only the topic proportion data and the forecast year are used as predictors. The forecast year is added to avoid spurious correlation as some topics may be correlated with the forecast year. I reject the null hypothesis of efficient forecasts for the construction forecast of the ifo (H4) and IfW (H0/H4). In these cases, I can show that the forecast reports contain additional information which could lead to improved prognosis for the same period. In the last four columns of Table 4, I additionally test for weak forecast efficiency. Weak forecast efficiency holds if preceding forecast errors cannot explain current errors. (Nordhaus 1987; Timmermann 2007). Therefore, the lagged forecast error is added, which should not improve the levels of significance. Weak forecast efficiency does not hold for the two-period CAP and the H0 EQIP prognosis of the ifo institute. Also, there is a reduction of the p-value in case of the ifo equipment forecast. It has to be noticed that the four-period construction forecast of the IfW institute turns insignificant after the inclusion of the lagged forecast error. However, it was barely significant in the previous configuration.
The left columns of Table 5 use the indicators and the forecast year as predictors, while the right columns also add the topic proportions and the lagged forecast error. With solely the indicators as predictors, the four-period GFCF forecasts of two institutes are proven to be inefficient. Also, I find multiple inefficient forecasts of machinery and equipment over all institutes. Without the usage of textual information, I fail to find the inefficiency of the construction forecasts. As anticipated, with the combined usage of indicators and topics as predictors my previous results are mostly merged. As the p-values measure the combined significance of all predictors, there are some new significant results at the

forecasts of CAP and EQIP. Simultaneously, a higher amount of predictors could lead to more predictors without explanatory power and therefore raise the p-values. As the combination of traditional indicators and textual information leads to the most significant results, I conclude a complementary relationship between both types of predictors. In the Tables 6 and 7, I vary the hyperparameters to test the robustness of my results. Table 6 uses all predictors and shows the results of 3 and 7 as *nodesize*. At a *nodesize* of 3, every terminal node could consist of tree observation, and therefore, the individual tree grows considerably deeper. Although the larger depth of the trees could lead to overfitting, good results are achieved with this setting, as the total number of significant observations stays constant. Turning the *nodesize* to 7 leads to trees with less depth and could improve the performance in the case of more variables without explanatory power, as Segal (2004) has shown. This configuration impairs my findings, as I cannot reject the null hypothesis for the H4 EQIP forecasts anymore. In Table 7, the parameter *mtry* is varied in the first columns to 6 ($\sqrt{p}$), which is the default value for classification trees, and in the last columns to 19, which equals $p/2$. A higher *mtry* leads to more similar and often correlated trees. In these trees, variables with moderate effect tend to have less impact as those with a high effect mask them. In random forests with lower values of *mtry* moderate variables are better exploited. However, as many trees are build based on suboptimal variables, the trees are performing worse on average (Probst et al. 2019). The variation of this parameter leads to mixed results. In some cases, the investigated forecast errors could be explained with several predictors of moderate effect (e.g. DIW and ifo H4 EQIP) as they lower their p-value with lower *mtry*. Other inefficiencies result from the small number of predictors with high explanatory power (e.g. ifo H4 CAP and IfW H0 BUILD) and therefore achieve lower p-values with higher values of *mtry*.
Figures 8 to 10 show the relevance of each predictor for the estimations with my default settings. I compute the mean squared error of each random forest model with withheld out-of-bag observations. Afterward, I permute a predictor and measure the percentual increase of the mean squared error (%IncMSE). Partial dependence plots are shown in Figures 11 to 12 for the three most impactful predictors of each at the 10% level of significance inefficient forecast. These show the predicted prognosis error for different values of the respective predictor with all other predictors fixed at their mean value. As every predictor has a non-linear relationship with the dependent variable, my random forest method is well suited for the revealing of these connections.
I could prove the H4 forecasts of GFCF to be inefficient, with those of the ifo strongly and those of the DIW and the IfW less significant. The %IncMSE plots show a strong effect from changes in stock market returns on the forecast error of ifo institute. As the investigation of the corresponding partial dependence plot reveals, the forecast errors rise to more than double their mean size if the stock market returns diminish. Aside from the stock market returns, I find several insufficient exploited indicators with a moderate effect on the H4 forecast errors of all institutes as the change of the business climate, the expected business climate, the money supply M1 and the inflow of industrial orders. As in my previous finding, I observe above average results in the case of a positive development of the indicators. With contraction or stagnation of these indicators, I observe a systematic overestimation of CAP.

Topic 16 ("GDP") is to be found as the second-biggest explanatory for the IfW and the biggest for the DIW forecast errors. With a topic proportion of around or below 0.08, the predicted forecast error lies beneath the mean forecast error. For higher proportion values of this topic, the predicted forecast errors rise. However, as the topic proportion exceeds 0.09 only in four years (1999, 2000, 2001, 2011), the big rise of the predicted forecast error has to be interpreted cautiously.
The H0 construction prognosis of the IfW is inefficient with high significance. The predictor with the highest %IncMSE is an indicator (consumer prices), whereby high inflation leads to an underestimation of construction forecasts. I find three topics with moderate explanatory power: Topic 21 (business expectations), Topic 17 (monetary policy) and Topic 16 (GDP). Each of these three topics addresses an essential determinant of the construction investment. Their respective partial dependence plots reveal that forecasts at topic proportions near zero are inferior. As low topic proportions could be interpreted as a lower emphasis on this topic, I explain the inefficient forecasts with the disregard of the determinants for construction investments. Therefore, my textual data implies a neglection of information for the creation of the prognosis.
I could prove the forecasts for investments in machinery and equipment inefficient in several cases because of underutilized indicators. The institutes tend to underestimate the impact of the decline of some economic indicators again. Specifically, the drop in stock market returns, business climate, term spread and industrial order inflow does not get appropriate integrated in the forecasts. Additionally, I find negative effects on forecast accuracy for a low Federal Fund rate and lower growth of money supply.
I use a subsample analysis to inspect the potential sensitivity of my results with respect to the large forecast errors that the research institutes made at the time of the German reunification and during the financial crisis of 2007/2008. In a subsample analysis, potential biases due to large forecast errors in the time of German reunification and the 2007/2008 financial crisis are addressed. Table 8 shows that the exclusion of the German reunification leads to fewer significant EQIP forecasts in both subsamples. As my previous results have shown, those forecasts inefficiencies emerge primarily from insufficient integration of downward indicators. Since most of the impactful indicators fell during the excluded years, it is plausible to find fewer significant results. Interestingly, I observe more inefficient construction forecasts, where the %IncMSEs are higher for the topics.

8 Conclusion

I used LDA in combination with word embeddings to process textual information of 584 forecast reports of 5 German forecast institutes from 1960 to 2017. My textual analysis shows a remarkable increase in the forecast report diversity over time. While a few big topics dominate the early forecast reports, the newer reports contain more specialized topics. For instance, I found two general policy topics whose topic proportions peak in the 1980s and subsequently partitioning into multiple smaller policy topics. Additionally, I could prove the use of topic models for the identification of the priorities of forecast-

ers. For example, there is a decline of the sectoral topic over the years and a rise of an immigration-labor market topic. Further, I could observe shifting prioritization in the discussion of economic issues. While the topic proportions of both investment topics are constant over time, forecasters consider investments less in the context of aggregate demand and more in the context of infrastructure. Although my applied method works sufficient over the long period, further research could work with a dynamic version of topic models with word embeddings and thus capture shifting views on economic issues in the same topic.

In the second part of my research, I integrated my topic proportions in random forest models to test the forecast efficiency of gross fixed capital formation, investment in construction and investment in machinery and equipment from 1970 to 2017 by three German forecast institutes. There are some cases in which textual information proves forecasts to be inefficient. In most cases, low topic proportions of relevant topics lead to inferior forecasts. With the inclusion of traditional predictors, substantially more inefficient forecasts are found. Especially downward indicators are often not utilized sufficiently. Even though the indicators have more impact on the prognosis errors in most cases, many topics with medium impact could be found. I suggest that textual data could be seen as an indicator of the information used by forecasters. Therefore, researchers could use textual data as a tool for the detection of inefficient forecasts, thus reveal insufficient exploited information and potentially improve future forecasts.

References

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016). "Measuring Economic Policy Uncertainty*". In: *The Quarterly Journal of Economics* 131.4, pp. 1593–1636. ISSN: 0033-5533. DOI: 10.1093/qje/qjw024.

Beckers, Benjamin, Konstantin A. Kholodilin, and Dirk Ulbricht (2017). "Reading between the Lines: Using Media to Improve German Inflation Forecasts". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.2970466.

Behrens, Christoph (2019). "A Nonparametric Evaluation of the Optimality of German Export and Import Growth Forecasts Under Flexible Loss". In: *Economies* 7.3, p. 93. DOI: 10.3390/economies7030093.

Behrens, Christoph, Christian Pierdzioch, and Marian Risse (2018a). "A test of the joint efficiency of macroeconomic forecasts using multivariate random forests". In: *Journal of Forecasting* 37.5, pp. 560–572. ISSN: 02776693. DOI: 10.1002/for.2520.

– (2018b). "Testing the optimality of inflation forecasts under flexible loss with random forests". In: *Economic Modelling* 72, pp. 270–277. ISSN: 02649993. DOI: 10.1016/j.econmod.2018.02.004.

– (2020). "Do German economic research institutes publish efficient growth and inflation forecasts? A Bayesian analysis". In: *Journal of Applied Statistics* 47.4, pp. 698–723. ISSN: 0266-4763.

Blei, David M. (2012). "Probabilistic topic models". In: *Communications of the ACM* 55.4, p. 77. ISSN: 00010782. DOI: 10.1145/2133806.2133826.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.

Bollen, Johan, Huina Mao, and Xiaojun Zeng (2011). "Twitter mood predicts the stock market". In: *Journal of Computational Science* 2.1, pp. 1–8. ISSN: 18777503. DOI: 10.1016/j.jocs.2010.12.007.

Choi, Hyunyoung and Varian Hal (2012). "Predicting the Present with Google Trends". In: *Economic Record* 88.1, pp. 2–9. ISSN: 00130249.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei (2019). *Topic Modeling in Embedding Spaces*.

Döpke, Jörg, Ulrich Fritsche, and Gabi Waldhof (2019). "Theories, Techniques and the Formation of German Business Cycle Forecasts". In: *Jahrbücher für Nationalökonomie und Statistik* 239.2, pp. 203–241. ISSN: 0021-4027.

Drechsel, Katja and Rolf Scheufele (2012). "The performance of short-term forecasts of the German economy before and during the 2008/2009 recession". In: *International Journal of Forecasting* 28.2, pp. 428–445. DOI: 10.1016/j.ijforecast.2011.04.003.

Evans, Jeffrey S. and Melanie A. Murphy (2018). *rfUtilities*. URL: https://cran.r-project.org/package=rfUtilities.

Feinerer, Ingo (2013). *Introduction to the tm Package Text Mining in R*. URL: http://cran.%20r-project.%20org/web/packages/tm/vignettes/tm.%20pdf.

Fritsche, Ulrich and Johannes Puckelwald (2018). *Deciphering professional forecasters' stories: Analyzing a corpus of textual predictions for the German economy*. Hamburg.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as Data". In: *Journal of Economic Literature* 57.3, pp. 535–574. ISSN: 0022-0515. DOI: 10.1257/jel.20181020.

Gentzkow, Matthew and Jesse M. Shapiro (2010). "What Drives Media Slant? Evidence From U.S. Daily Newspapers". In: *Econometrica* 78.1, pp. 35–71. ISSN: 0012-9682. DOI: 10.3982/ECTA7195.

Greenstein, Shane M., Grace Gu, and Feng Zhu (2017). *Ideology and Composition Among an Online Crowd: Evidence from Wikipedians*.

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101 (suppl. 1).

Hassan, Tarek A. et al. (2019). "Firm-Level Political Risk: Measurement and Effects*". In: *The Quarterly Journal of Economics* 134.4, pp. 2135–2202. ISSN: 0033-5533. DOI: 10.1093/qje/qjz021.

Hastie, Trevor, Jerome H. Friedman, and Robert Tibshirani (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York.

Heinrich, Gregor (2009). *Heinrich, Gregor. Parameter estimation for text analysis*.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103. New York, NY: Springer New York.

Liaw, Andy and Matthew Wiener (2002). "Classification and Regression by randomForest". In: *R News* 2/3.

Lucca, David and Francesco Trebbi (2009). *Measuring central bank communication: an automated approach with application to FOMC statements.*

Mathy, Gabriel and Herman Stekler (2017). "Expectations and forecasting during the Great Depression: Real-time evidence from the business press". In: *Journal of Macroeconomics* 53, pp. 1–15. ISSN: 01640704. DOI: 10.1016/j.jmacro.2017.05.006.

Mimno, David et al. (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing.*

Nordhaus, William D. (1987). "Forecast efficiency: concepts and applications". In: *The Review of Economics and Statistics.*

Panigrahi, Abhishek, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya, eds. (2019). *Word2Sense: Sparse Interpretable Word Embeddings.* Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics.

Probst, Philipp, Bernd Bischl, and Anne-Laure Boulesteix (2018). *Tunability: Importance of Hyperparameters of Machine Learning Algorithms.*

Probst, Philipp and Anne-Laure Boulesteix (2018). "To Tune or Not to Tune the Number of Trees in Random Forest". In: *Journal of Machine Learning Research* 18.

Probst, Philipp et al. (2019). "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3, p. 281. ISSN: 1942-4787.

R Core Team (2019). *R: A language and environment for statistical computing.* Vienna, Austria. URL: https://www.R-project.org/..

Scott, Steven L. and Hal R. Varian (2014). "Predicting the Present with Bayesian Structural Time Series". In: *International Journal of Mathematical Modelling and Numerical Optimisation* 5.1-2. DOI: 10.2139/ssrn.2304426.

– (2015). "Bayesian Variable Selection for Nowcasting Economic Time Series". In: *Economic analysis of the digital economy.* Ed. by Avi Goldfarb et al. National Bureau of Economic Research conference report. Chicago and London: The University of Chicago Press.

Segal, Mark (2004). "Machine learning benchmarks and randon forest regression". In:

Sinclair, Tara M. and H. O. Stekler (2013). "Examining the quality of early GDP component estimates". In: *International Journal of Forecasting* 29.4, pp. 736–750.

Stekler, Herman and Hilary Symington (2016). "Evaluating qualitative forecasts: The FOMC minutes, 2006–2010". In: *International Journal of Forecasting* 32.2, pp. 559–570. DOI: 10.1016/j.ijforecast.2015.02.003.

Tetlock, Paul C. (2007). "Giving content to investor sentiment: The role of media in the stock market". In: *The Journal of Finance* 62.3, pp. 1139–1168. ISSN: 00221082. DOI: 10.1111/j.1540-6261.2007.01232.x.

Timmermann, Allan (2007). "An evaluation of the World Economic Outlook forecasts". In: *IMF Staff Papers* 54.1.

9 Tables and Figures

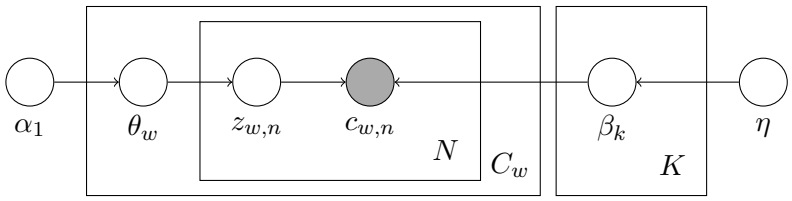

Figure 1: Our generative Model for the co-occurrence matrix. The plates denote replication. The observed node are shaded grey.

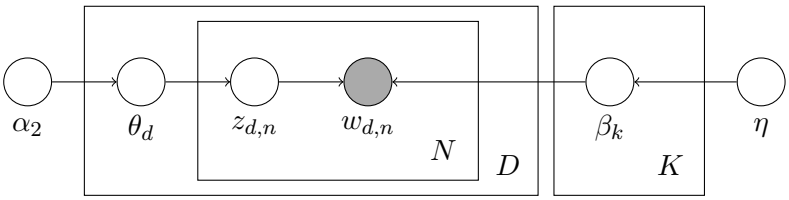

Figure 2: Standard LDA model. The plates denote replication. The observed node are shaded grey. Own representation based on Blei 2012.

Table 1: Topic labels

Topic	Coherence	Diversity	Labels	Top Ten Terms
t_1	0.31	0.56	context words	undertake/company[1], extensive, measures, case, problem, increasing, federal republic, let, give, quickly
t_2	0.38	0.28	investments	demand, production, building investment, equipment investments, investment, residential construction , cyclical, abroad, industry, company
t_3	0.17	0.72	self-references	part, institute, view, period, Gemeinschaftsdiagnose, Weltwirtschaft, importance, economic research, prices, look
t_4	0.35	0.28	government spending	expanses, public, state, euro, federation, gross domestic product, investments, deficit, measures , revenues
t_5	0.14	0.84	context words	international, defect, trends, get, risk, state, seem, insofar as, related, cancellation
t_6	0.35	0.60	inflation	price, euro, Deutsche Mark, US-dollar, consumer price, dollar, price buoyancy, price increase, price rise, revaluation
t_7	0.32	0.72	unemployment benefits/immigration	persons, unemployment benefit, line, refugee, population, reserve, measures, unemployed, labor market
t_8	0.36	0.68	wages	wages, employees , tariff, salaries, wage, standard wage, service, profit, employ, working time
t_9	0.34	0.00	overview	production, Germany, business cycle, company, demand, gross domestic product, export, economy, cyclical, expansion
t_10	0.33	0.64	sectoral	business, industry, production, import, trade, retail, areas, capital goods, services, processing
t_11	0.32	0.68	statistical	forecast, institute, federal office, statistical, assumptions, basis, data, gross domestic product, statistical, present
t_12	0.27	0.84	recession	recession, early, winter, weather conditions, upswing, financial market, crisis, slump, uncertainty, financial crisis
t_13	0.40	0.44	context words	pace, accelerated, expand, import, price increase, extended , abundant, capacity utilization, expected, relevant
t_14	0.24	0.64	general policy	measures, federal government, objective, financial policy, frame, policy, stability, federal bank, economic policy, fiscally
t_15	0.40	0.32	international trade	German, export, competitiveness , exports, countries, euro area, European, Germany, economic activity, states
t_16	0.33	0.64	GDP	gross domestic product, increase , growth, lie, growth rates, comparison, production, abundant, rate, rate of increase
t_17	0.31	0.64	monetary policy	Bundesbank, interest, banks, monetary policy, money supply , long-term, credit, company, Germany, European
t_18	0.36	0.40	employment	persons, unemployment, employment, unemployed , labor market, employee, work, working population, work time, short time
t_19	0.29	0.60	current account	term, delimitation, economically, current account, deficit, trade, account , surplus, improve, net export
t_20	0.35	0.16	general policy	company , monetary policy, financial policy, measures, economic activity, Germany, Federal Bank, effect, prognosis, growth
t_21	0.33	0.76	business expectations	company, business expectations, business situation, propitious , improved, capacities, expectations, improvement, current , more favorable
t_22	0.41	0.72	taxes/social insurances	statutory, contribution rate , increase , taxes , raise, increased, pension insurance, revenue, introduction , health insurance
t_23	0.18	0.72	investments	investments , promotion, apartments, company , Treuhandanstalt, residential construction, expire, facilities/assets[2], infrastructure, program
t_24	0.37	0.28	household income	private, household, income, consumption, public, private, company, investments, demand, expanses

Notes: Grey underlined topics cannot be assigned to a specific economic subject. [1] The German word "unternehmen" could be translated into "undertake" or "company".
[2] Although likely both meanings are included here, we assume a significant share of the former meaning. The word "anlagen" could be translated into "facilities" or "assets".

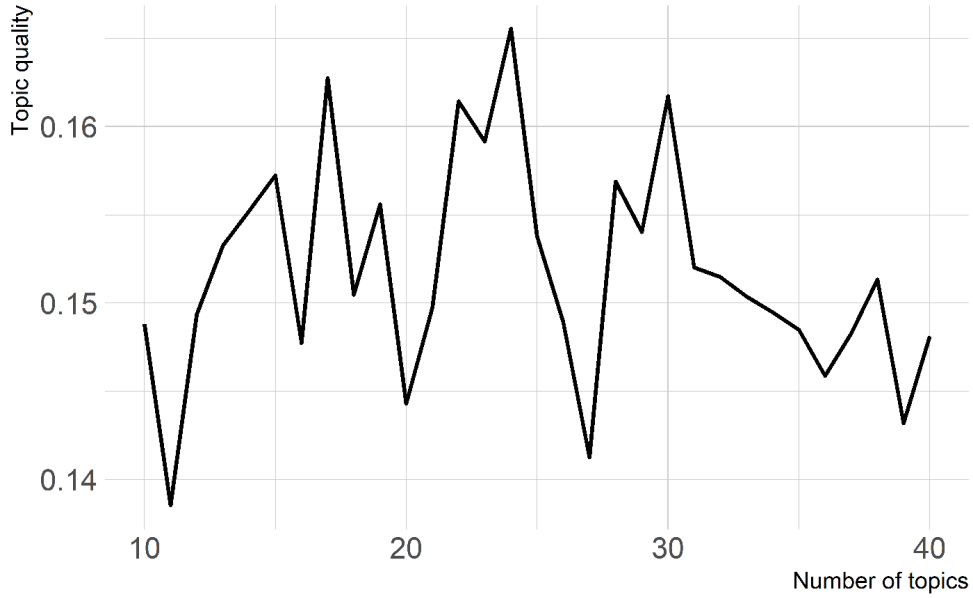

Figure 3: Topic quality for different number of topics.

Table 2: Descriptive statistics of forecast errors.

Institute	Horizon	*N*	MEAN	SD	*N*	MEAN	SD	*N*	MEAN	SD
			CAP			***BUILD***			***EQIP***	
DIW	**H0**	48	0.43	0.39	48	0.55	0.45	42	0.75	0.66
	H2	37	1.67	1.02	37	1.76	1.26	31	2.59	1.70
	H4	49	2.25	1.68	49	2.04	1.64	43	3.27	2.90
	H6	36	3.36	2.70	36	2.77	1.92	30	4.98	5.41
ifo	**H0**	28	0.56	0.95	44	0.72	1.17	40	0.68	0.64
	H2	24	1.46	1.07	42	1.70	1.17	40	2.56	1.87
	H4	30	2.19	1.74	45	2.31	1.58	41	3.48	2.93
	H6	25	3.31	2.56	38	2.74	1.86	36	4.49	4.45
IfW	**H0**	27	0.40	0.51	46	0.90	0.93	46	1.08	2.09
	H2	25	1.89	1.38	41	1.68	1.16	41	3.17	2.52
	H4	28	2.28	1.81	46	2.12	1.44	46	4.12	3.72
	H6	25	4.04	2.75	41	2.57	1.91	41	5.56	4.93

Notes: ***N***: Number of observations, **MEAN** Arithmetic mean, **SD**: Standard deviation.

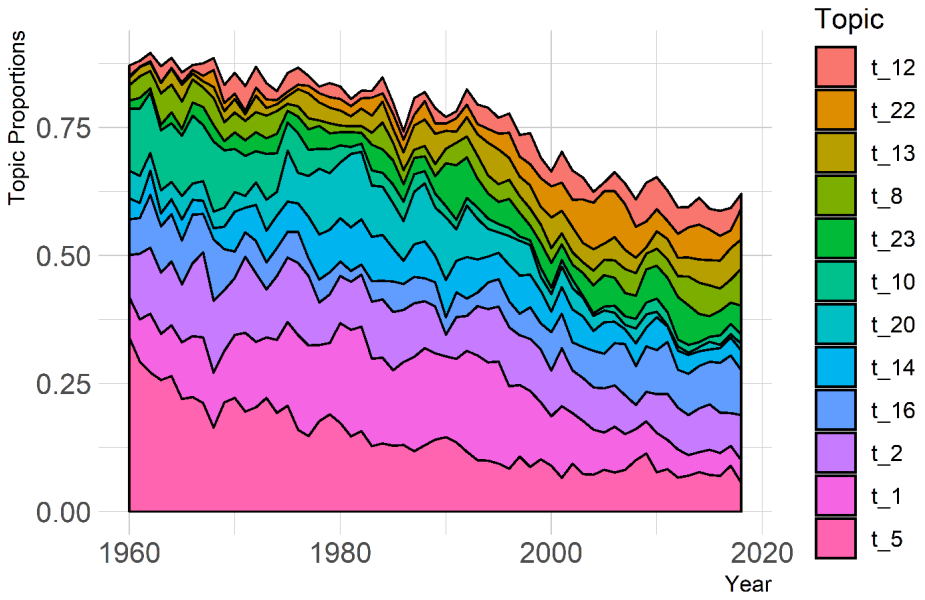

Figure 4: Topic proportion of the 12 most frequent topics.

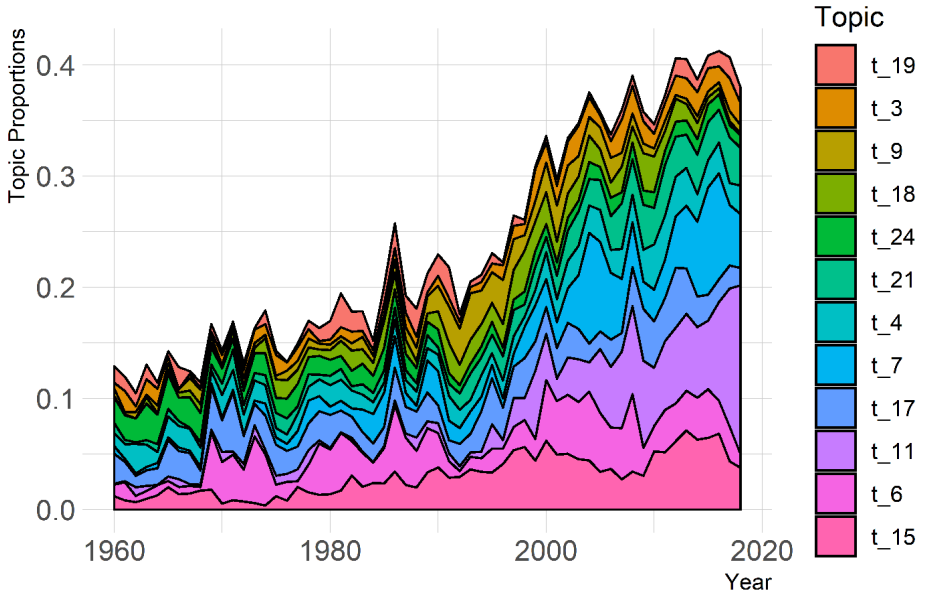

Figure 5: Topic proportion of the 12 least frequent topics.

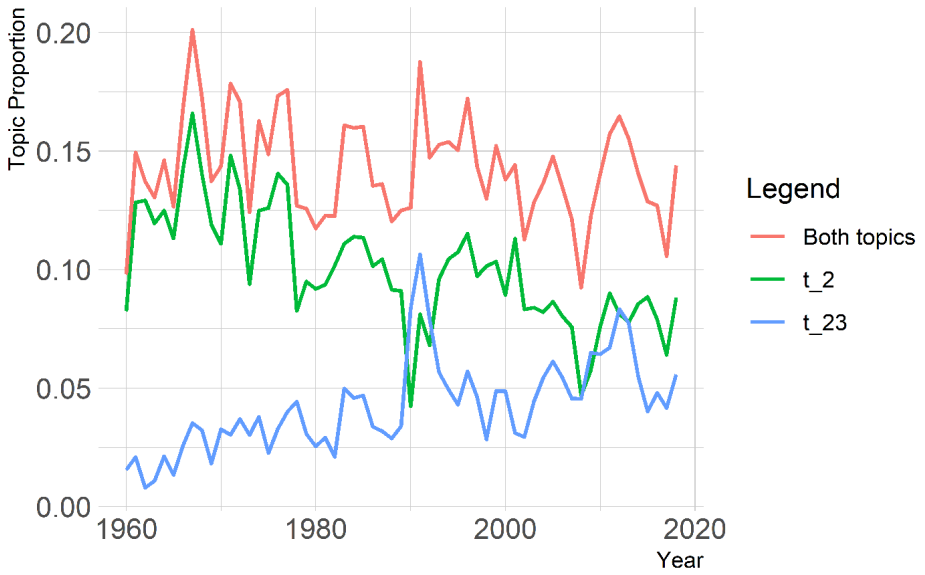

Figure 6: Topic proportion of investment related topics.

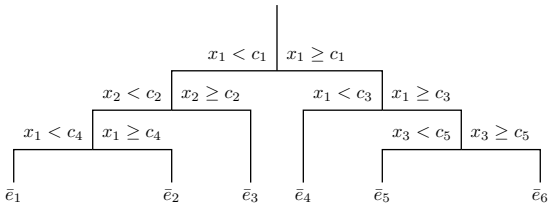

Figure 7: Example of simple regression tree. $x_1, .., x_3$ denote participating predictors, $c_1, .., c_5$ are used as splitting points and $\bar{e}_1, .., \bar{e}_6$ are the predicted outcomes. Own representation based on Behrens et al. (2018a).

Table 3: Indicators

Predictors	Acronym	ln	Lag	Description	Source
3 months interest rate	RK	N	0	Monthly average of 3 months money market rate	BUBA
1 month interest rate	RK1	N	0	Monthly average of 1 month money market rate	BUBA
discount rate	DR	N	1	discount rate at the end of the month	BUBA
Term spread	SPREAD	N	0	Monthly average of the yield on debt securities with a maturity of more than 3 years minus RK	BUBA
Order inflow	ORDER	Y	1	Year-on-year rate of the change of industrial orders; calendar and seasonally adjusted	BUBA
Climate	CLIMATE	N	0	Monthly ifo business tendency survey; seasonally adjusted	FRED
Climate (expectations)	CLIMATE_EXP	N	0	Monthly ifo business tendency survey; situation in 6 months; seasonally adjusted	FRED
Consumer Prices	INF	Y	0	Year-on-year rate of change of the monthly consumer price index; calendar and seasonally adjusted	BUBA
Stock market returns	STOCK	Y	0	Year-on-year returns on the share-price index	OECD
Industrial production	PRODUCTION	Y	1	Year-on-year growth rate of industrial production	BUBA
Dollar exchange rate	DOLLAR	Y	1	Year-on-year change of the exchange rate of the dollar vis-à-vis the euro	FRED
U.S. production	US_PROD	Y	1	Year-on-year growth rate of U.S. industrial production	OECD
Oil price	OIL	Y	0	Year-on-year change of the monthly crude oil price (WTI) Oil in Dollar	FRED
OECD leading	OECD_LEADING	N	2	Monthly normalized OECD composite leading indicator	FRED
Real effective exchange rate	REER	Y	1	Year-on-year rate of change of the monthly narrow effective exchange rate; CPI based	FRED
U.S. Federal Funds Rate	RK_US	N	0	Effective monthly U.S. federal funds rate	FRED
Money supply M1	M1	Y	1	Monthly change of money supply M1; seasonally adjusted	FRED/ BUBA

Notes: All indicators are for Germany, if not specified otherwise; BUBA - German Central Bank; FRED - Federal Reserve Bank of St.Louis; OECD -Organisation for Economic Co-operation and Development; ln: natural logarithmic transformation, Y - yes, N - no; Lag: Publication lags in months added where necessary.

Table 4: Topic proportions as predictors

Horizon	**H0**	**H2**	**H4**	**H6**	**H0**	**H2**	**H4**	**H6**
Predictors	*Topic models*				*Topic models & lags*			
	Gross fixed capital formation							
DIW	0.566	0.209	0.601	0.569	0.738	0.369	0.542	0.661
ifo	0.192	0.104	0.335	0.766	0.158	**0.064**	0.486	0.763
IfW	0.676	0.362	0.216	0.573	0.411	0.656	0.116	0.841
	Construction							
DIW	0.408	0.843	0.534	0.369	0.487	0.891	0.653	0.444
ifo	0.102	0.193	**0.071**	0.349	0.12	0.288	**0.053**	0.548
IfW	**0.001**	0.116	**0.079**	0.449	**0.002**	0.119	0.11	0.176
	Machinery and equipment							
DIW	0.777	0.535	0.688	0.973	0.653	0.678	0.26	0.845
ifo	0.162	0.429	0.556	0.654	**0.007**	0.367	0.532	0.569
IfW	0.852	0.264	0.692	0.612	0.83	0.176	0.759	0.446

Notes: Reported p-values are obtained by a permutation test with 1000 replications. Bold numbers indicate significance at the 10%-level.

Table 5: Indicators and topic porportions as predictors

Horizon	**H0**	**H2**	**H4**	**H6**	**H0**	**H2**	**H4**	**H6**
Predictors			*Indicators*				*Topic models & indicators*	
				Gross fixed capital formation				
DIW	0.355	0.623	**0.012**	0.291	0.516	0.305	0.109	0.284
ifo	0.749	0.448	**0.002**	0.201	0.629	0.124	**0.005**	0.384
IfW	0.918	0.707	0.129	0.458	0.798	0.553	**0.039**	0.712
				Construction				
DIW	0.279	0.99	0.525	0.873	0.645	0.984	0.786	0.851
ifo	0.996	0.698	0.285	0.826	0.961	0.632	0.171	0.549
IfW	0.576	0.69	0.54	0.769	**0.03**	0.365	0.136	0.625
				Machinery and equipment				
DIW	0.348	0.172	**0.051**	0.593	0.377	0.212	**0.073**	0.715
ifo	**0.006**	**0.024**	0.105	0.351	**0.015**	**0.073**	**0.08**	0.517
IfW	0.641	0.487	**0.018**	0.153	0.695	0.101	**0.099**	0.256

Notes: Reported p-values are obtained by a permutation test with 1000 replications. Bold numbers indicate significance at the 10%-level.

Table 6: Variation of *nodesize*

Horizon	**H0**	**H2**	**H4**	**H6**	**H0**	**H2**	**H4**	**H6**
nodesize			***3***				***7***	
				Gross fixed capital formation				
DIW	0.528	0.29	**0.08**	0.275	0.485	0.359	0.12	0.306
ifo	0.884	0.172	**0.003**	0.349	0.714	0.11	**0.002**	0.386
IfW	0.55	0.668	**0.031**	0.9	0.678	0.552	**0.061**	0.882
				Construction				
DIW	0.356	0.988	0.694	0.855	0.498	0.98	0.795	0.871
ifo	0.962	0.507	0.178	0.591	0.975	0.437	0.16	0.506
IfW	**0.048**	0.348	0.144	0.578	**0.044**	0.274	0.146	0.769
				Machinery and equipment				
DIW	0.442	0.152	**0.069**	0.789	0.569	0.151	0.11	0.728
ifo	**0.018**	**0.071**	0.132	0.53	**0.012**	**0.055**	0.135	0.408
IfW	0.832	**0.074**	**0.096**	0.201	0.795	**0.094**	0.121	0.148

Notes: Reported p-values are obtained by a permutation test with 1000 replications. Bold numbers indicate significance at the 10%-level.

Table 7: Variation of *mtry*

Horizon	**H0**	**H2**	**H4**	**H6**	**H0**	**H2**	**H4**	**H6**
mtry			***6***				***19***	
				Gross fixed capital formation				
DIW	0.531	0.267	**0.099**	0.357	0.637	0.293	0.181	0.253
ifo	0.456	0.111	**0.025**	0.35	0.921	0.102	**0.001**	0.459
IfW	0.591	0.667	**0.044**	0.646	0.762	0.668	**0.052**	0.942
				Construction				
DIW	0.424	0.982	0.662	0.744	0.675	0.996	0.816	0.894
ifo	0.706	0.539	0.13	0.517	0.993	0.679	0.181	0.574
IfW	**0.087**	0.393	0.107	0.59	**0.018**	0.3	0.186	0.579
				Machinery and equipment				
DIW	0.319	0.155	**0.072**	0.64	0.531	**0.086**	0.111	0.781
ifo	**0.01**	**0.084**	**0.085**	0.371	**0.015**	**0.045**	0.18	0.482
IfW	0.525	**0.085**	0.149	0.14	0.856	0.109	0.128	0.291

Notes: Reported p-values are obtained by a permutation test with 1000 replications. Bold numbers indicate significance at the 10%-level.

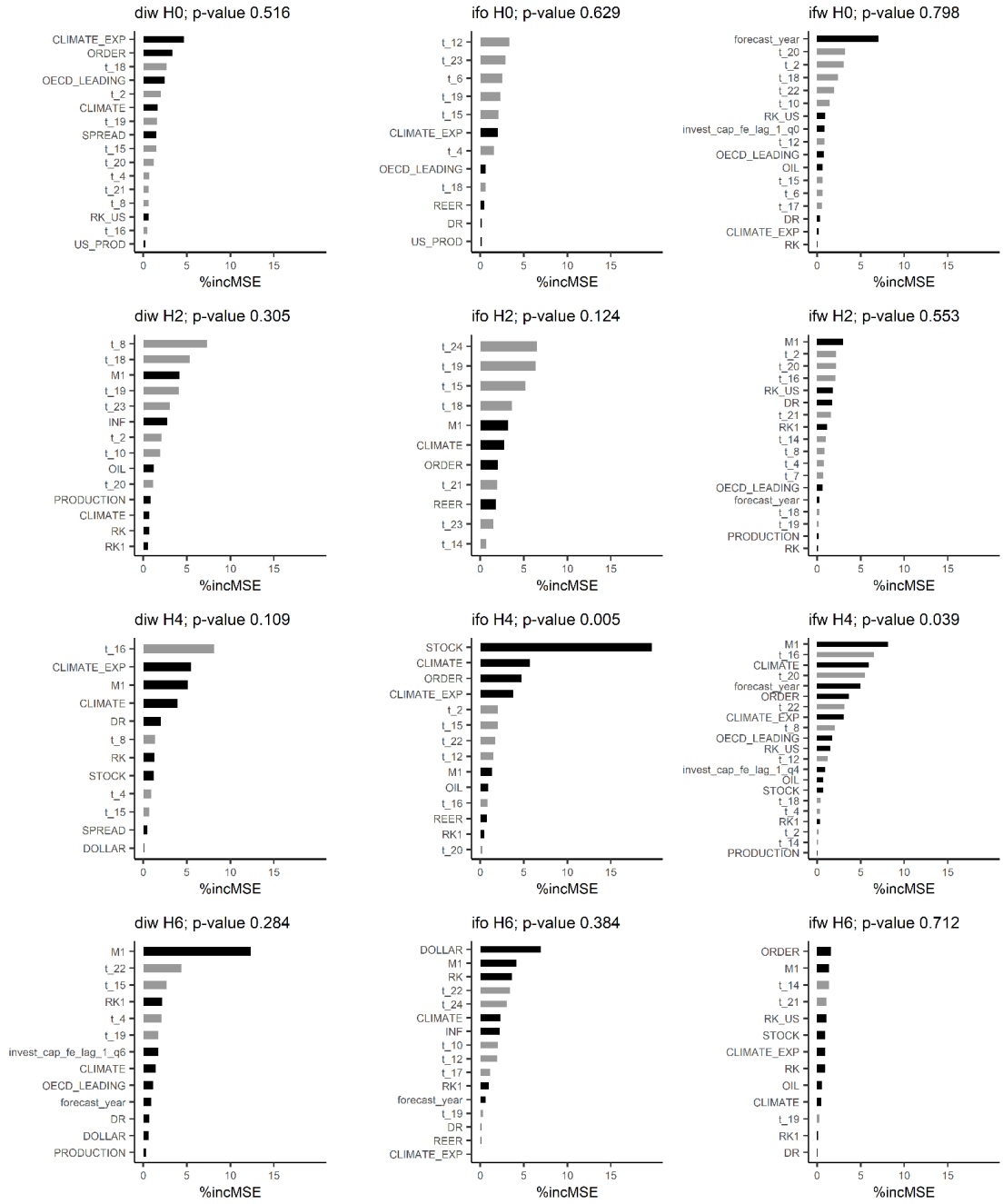

Figure 8: Plots of %IncMSE for random forest evaluating gross fixed capital formation forecast - All indicators (black) and topic proportions (grey) are used while only reporting predictors with a positive %IncMSE value. I use the standard settings of the literature.

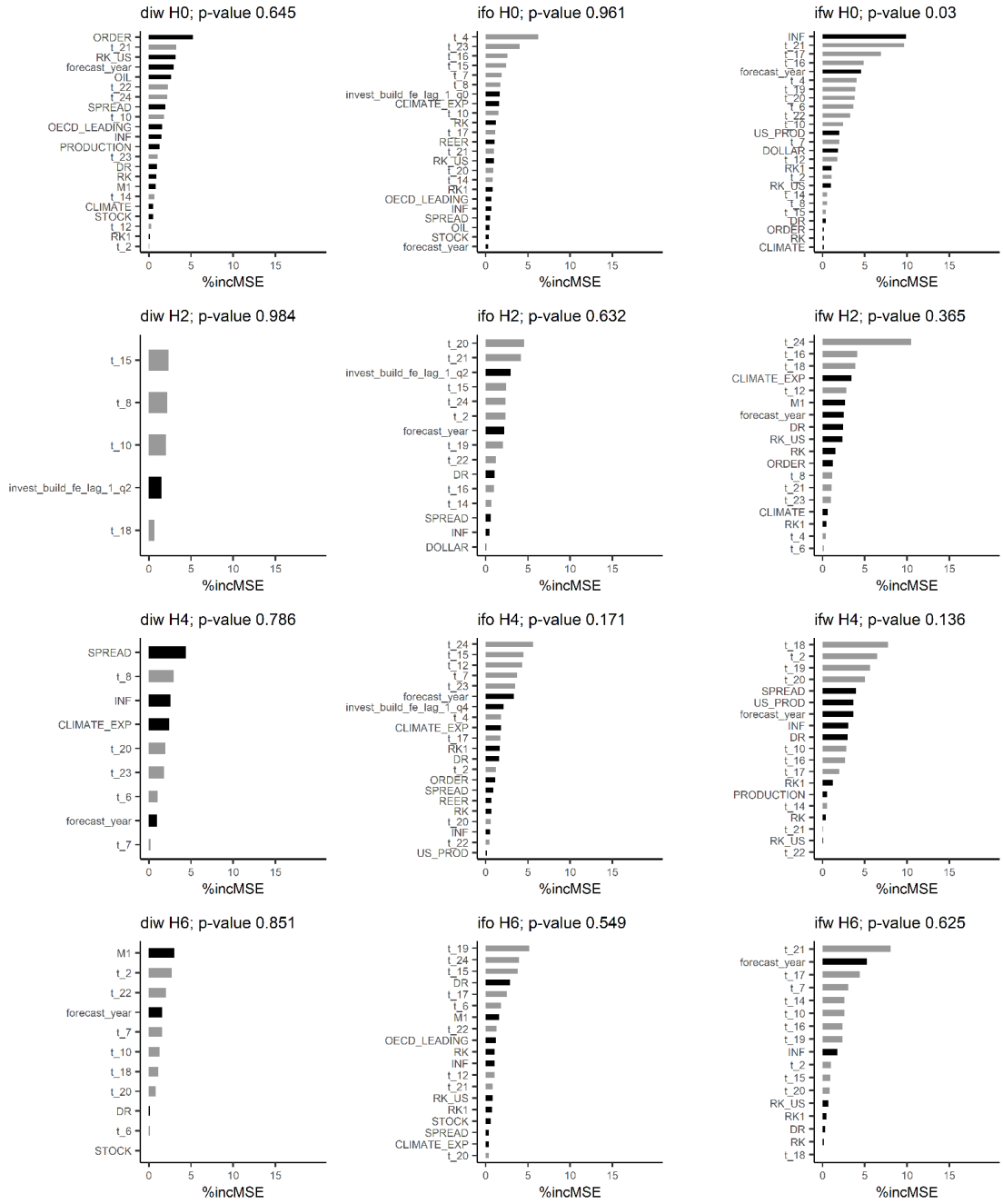

Figure 9: Plots of %IncMSE for random forest evaluating construction forecast - All indicators (black) and topic proportions (grey) are used while only reporting predictors with a positive %IncMSE value. I use the standard settings of the literature.

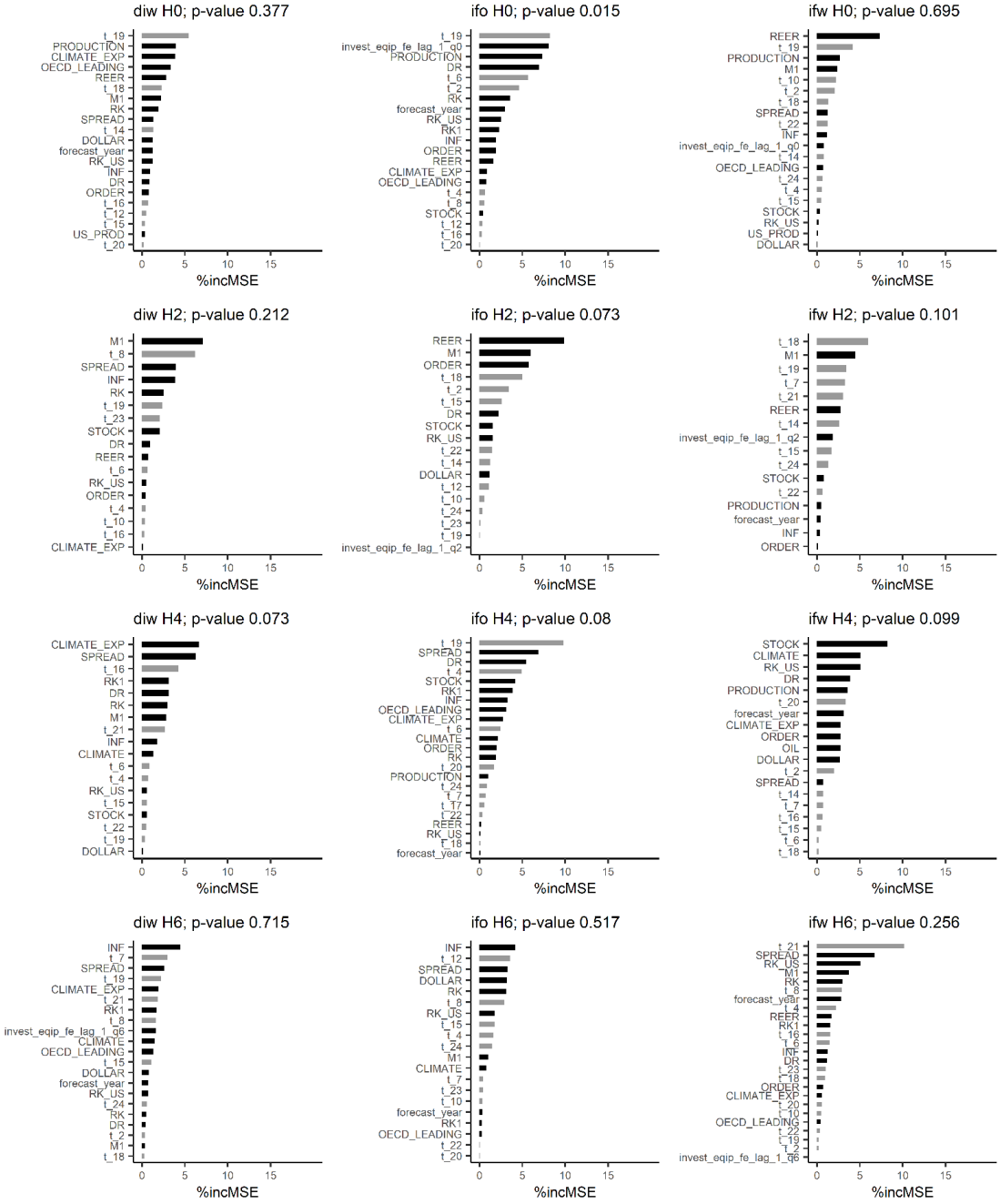

Figure 10: Plots of %IncMSE for random forest evaluating forecast in machinery and equipment - All indicators (black) and topic proportions (grey) are used while only reporting predictors with a positive %IncMSE value. I use the standard settings of the literature.

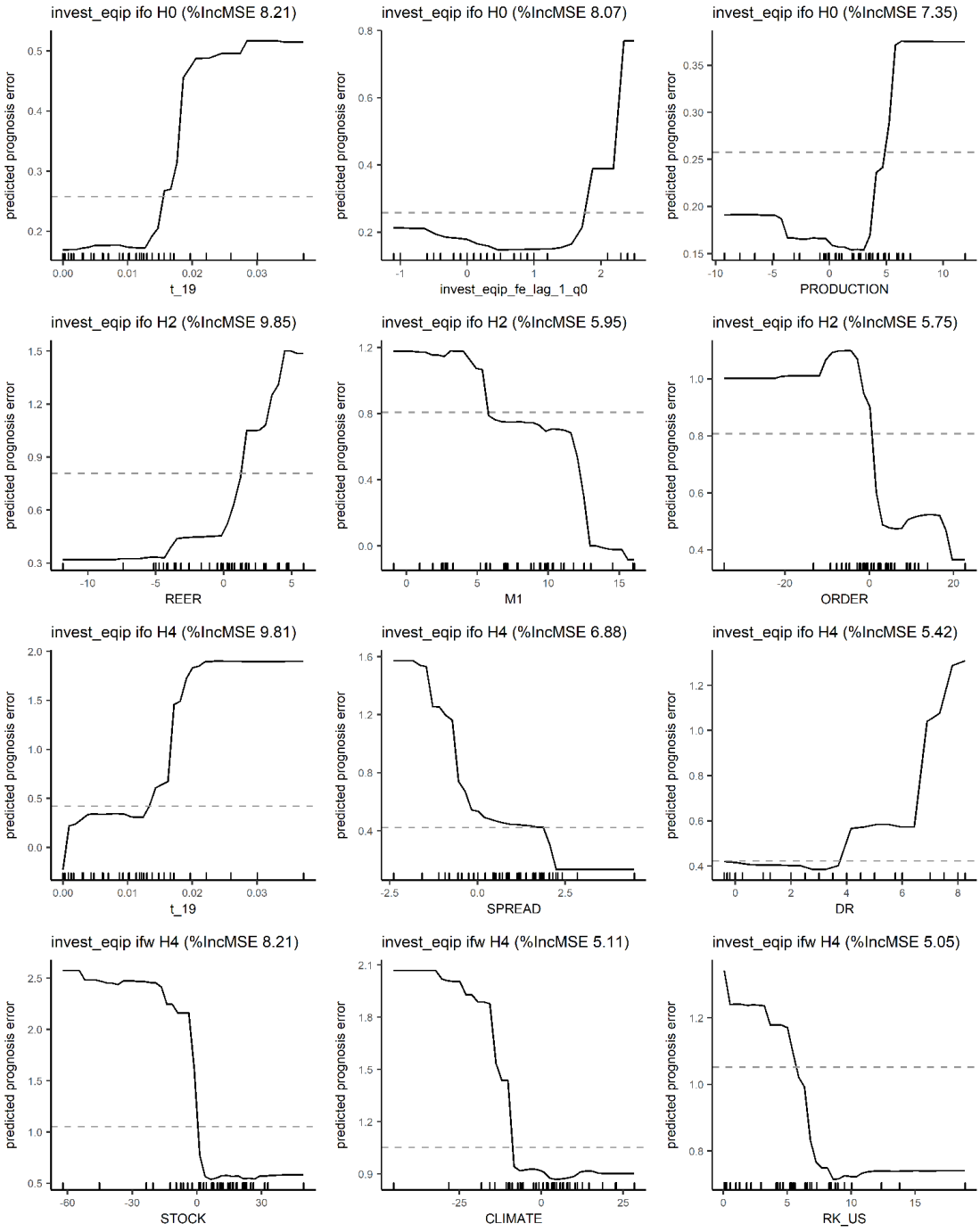

Figure 11: Partial dependence plots for the three most relevant predictors of inefficient forecasts. Rugs represent observations. Dashed line shows the mean prognosis error of the respective time series.

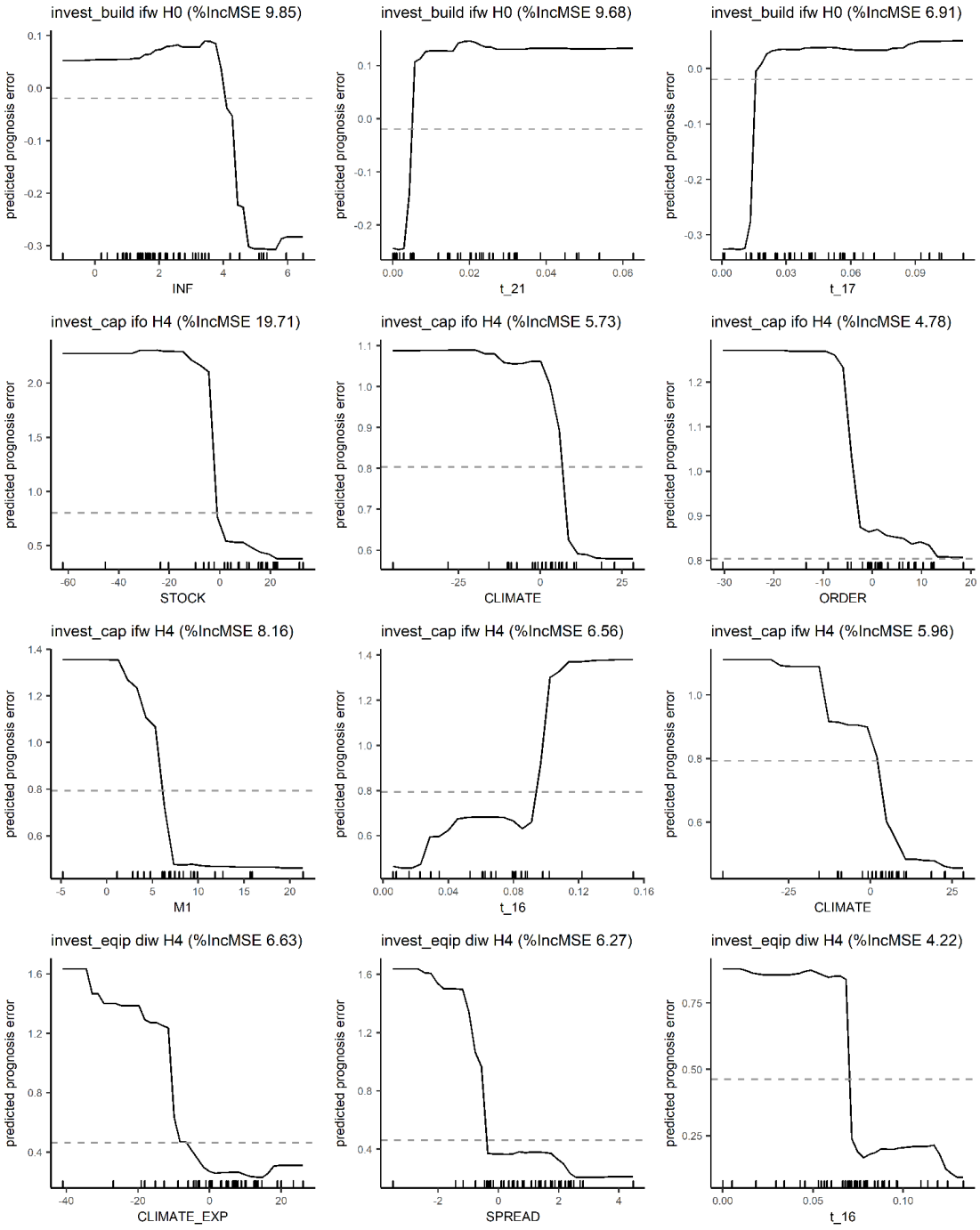

Figure 12: Partial dependence plots for the three most relevant predictors of inefficient forecasts. Rugs represent observations. Dashed line shows the mean prognosis error of the respective time series.

Table 8: Subsample analysis

Horizon	**H0**	**H2**	**H4**	**H6**	**H0**	**H2**	**H4**	**H6**
Predictors	*without German unification years*				*without financial crisis*			
	Gross fixed capital formation							
DIW	0.534	0.137	0.264	0.413	0.438	0.294	0.134	0.322
ifo	**0.06**	0.159	**0.013**	0.36	0.5	0.152	**0.015**	0.138
IfW	0.695	0.568	**0.041**	0.876	0.614	0.767	**0.051**	0.882
	Construction							
DIW	0.793	0.961	0.853	0.937	0.344	0.99	0.713	0.769
ifo	**0.092**	0.611	0.134	0.294	0.964	0.538	0.185	0.371
IfW	**0.012**	0.379	**0.087**	0.407	**0.02**	0.516	0.152	0.384
	Machinery and equipment							
DIW	0.566	0.238	0.189	0.562	0.409	0.249	0.119	0.403
ifo	0.108	0.169	0.239	0.682	**0.006**	0.104	**0.069**	0.201
IfW	0.262	**0.087**	0.198	0.377	0.237	**0.07**	0.372	0.256

Notes: Reported p-values are obtained by a permutation test with 1000 replications. Bold numbers indicate significance at the 10%-level. Subsamples are without 1992-1993 or 2007-2008.